Computación y programación en R: Tema 4

David V. Conesa Guillén

GEeitEma

Valencia Bayesian Research Group

Dept. d'Estadística i Investigació Operativa

Universitat de València

Tema 4: Análisis de datos con R

En este tema:

1.- Introducción a la modelización estadística.

2.- Modelos básicos con datos normales.

3.- Análisis de la Varianza.

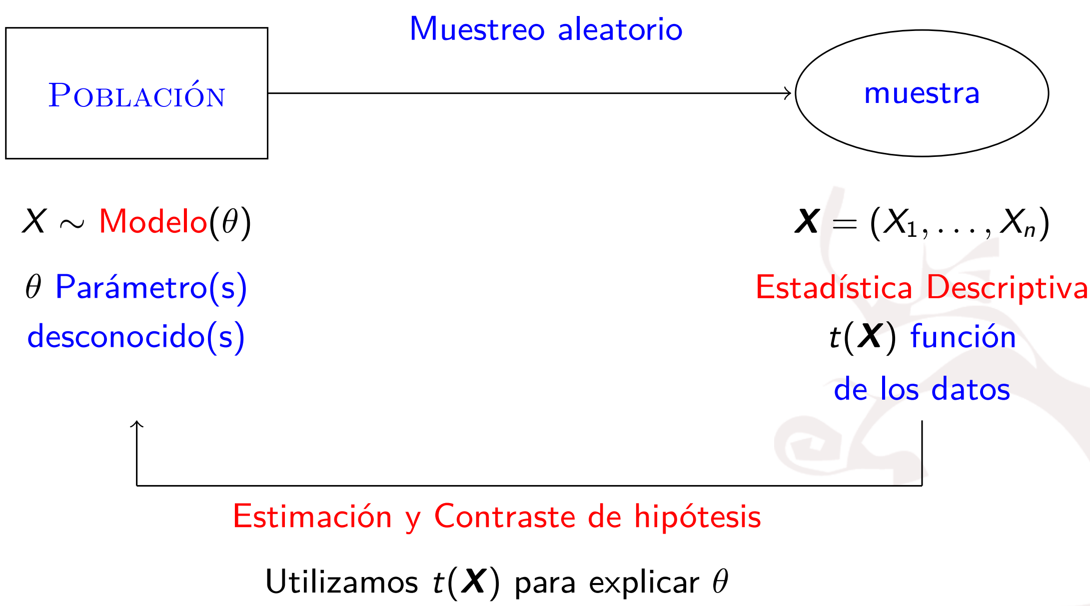
4.- Comparaciones.

5.- ANOVA de un factor con R.

Sección 1 | Introducción a la Modelización Estadística

Introducción a la modelización estadística.

- Los datos obtenidos cuando realizamos cualquier experimento presentan variabilidad y la Estadística nos permite analizar los datos que exhiben variabilidad.
- En ese proceso podemos distinguir entre Modelización estadística; Estadística descriptiva, e Inferencia estadística. Gráficamente:



Modelización en la Estadística.

- En general, un modelo es una representación en pequeña escala de la realidad.
- La Estadística nos permite incorporar la variabilidad presente en la vida real en nuestros modelos a través de la aleatoriedad.
- Aun así: "esencialmente, todos los modelos son incorrectos, pero algunos son útiles" (Box, 1987).
- Los modelos estadísticos son la base en la que se sustentan la mayoría de las técnicas de análisis de datos habituales.
- "La formulación del problema es más esencial que su propia solución, que puede ser simplemente una habilidad matemática o experimental" (Albert Einstein).

Formulación de un problema.

Para formular adecuadamente un problema:

- Entender el background físico del problema. Los estadísticos suelen trabajar de manera conjunta con especialistas de diversas áreas y necesitan conocer lo básico del área del problema.
- Comprender claramente el objetivo. No es una tarea fácil concretarlo.
- Tener claro lo que quiere la persona que nos ha consultado. No tiene sentido hacer un análisis más complejo si su interés es menor.
- ¡Poner el problema en términos estadísticos! Éste es el paso clave en el que a veces se cometen errores irreparables. Pero una vez está en términos estadísticos, la solución suele ser rutina.
- Pero cuidado, que tengamos un modelo que lea y procese los datos no es suficiente si no somos capaces de extraer conclusiones.

Comentarios sobre la toma de datos.

En el proceso de modelización es crucial el entender como se han tomado o como se van a tomar los datos.

- ¿Los datos son observacionales o experimentales?
- Es decir: ¿provienen de una muestra o encuesta convencional (recogidas por mera observación si control sobre las condiciones) o han sido obtenidos como resultado de un diseño experimental (con control sobre las condiciones)?
- Las conclusiones dependerán de que tipo son: en el primer caso podremos hablar de asociación entre los datos, mientras que en el segundo caso podremos hablar de una relación causa-efecto.
- ¿Hay información o datos que no hemos observado?
- ¿Hay valores faltantes?
- ¿Cómo se han codificado los datos?
- ¿Cuales son las unidades de medida? Redondeo.
- ¿Hay errores en la entrada de datos? Conviene realizar análisis previos.

Modelización: tipos de variables.

- Cuando modelizamos un problema real:
  - ¿qué queremos explicar?
  - y ¿en base a qué?
- Esto nos clasifica las variables en:
  - Variables respuesta: las que queremos explicar
  - Variables explicativas (o independientes, o predictoras): las que nos sirven para explicar las variables respuesta
- Las variables también se clasifican por su tipo de atributo:
  - Cualitativo $\longrightarrow$ Intrínsecamente no tiene carácter numérico (categórica)
    - ⋆ Nominal (sin orden entre los valores): Sexo
    - ⋆ Ordinal (con valores ordenados): Nivel de estudios
  - Cuantitativo $\longrightarrow$ Intrínsecamente numérico
    - ⋆ Discreto (cantidad finita o numerable de valores): Número de hijos
    - ⋆ Continuo (valores en toda la recta real): Altura

Modelos estadísticos.

- La mayoría de los modelos estadísticos tienen una estructura del tipo:
  - Variable respuesta que se quiere explicar.
  - Una componente sistemática que contiene la información "general" del sistema bajo estudio, y que se expresa como una combinación de variables explicativas en forma de ecuación paramétrica. Indica pues como afectan las explicativas a la respuesta.
  - Una componente aleatoria que refleja la variabilidad intrínseca en cada situación (en cada dato) particular.
- Dependiendo del tipo de variable, las explicativas son:
  - Cualitativo $\longrightarrow$ Factores (con sus correspondientes "niveles")
    - ★ Efectos fijos (si los niveles del factor están prefijados de antemano: Sexo)
    - ★ Efectos aleatorios (si los niveles del factor son una muestra aleatoria de los posibles niveles de dicho factor: persona)
  - Cuantitativo $\longrightarrow$ Covariables

Modelos estadísticos paramétricos.

- La mayoría de las veces la componente sistemática viene expresada como una combinación lineal (pero puede ser también no lineal).
- Una manera de incluir la aleatoriedad en la componente aleatoria es asignando una distribución de probabilidad a la variable respuesta.
- Si la variable respuesta es normal (la mayoría de los casos) y la relación lineal nos encontramos ante los modelos lineales.
- Ejemplo: explicar el peso de una persona por su altura y su edad.

$$\mu_Y = \beta_0 + \beta_1 X^{(1)} + \beta_2 X^{(2)}$$

- Esta explicación no es tan clara si la variable es discreta o cuantitativa pero la observamos categorizada.
- La variable respuesta no tiene porqué ser normal, podría ser binomial, Bernoulli, gamma, Poisson, etc.
- En cualquier caso, todos los modelos siempre vienen expresados en función de parámetros, siendo la inferencia sobre ellos (estimación y contraste de hipótesis) nuestro objetivo final.

Revisión elementos básicos inferencia

Estimación

- Utilizamos la estimación puntual de un parámetro para tener una primera aproximación sobre su valor.
- En la mayoría de los modelos se conoce el mejor estimador y un Intervalo de confianza para los parámetros de interés.

Elementos de un contraste

- Datos (obtenidos de forma muy diversa)
- Hipótesis nula (H_0)
- Hipótesis alternativa (H_A)
- Estadístico de contraste T (y su distribución bajo H_0)
- Valor observado del est. de contraste: t
- P-valor: Prob. si H_0 es cierta de que el valor de T sea más extremo que t en la dirección de la hip. alternativa

Revisión elementos básicos inferencia (y 2)

Contrastes Paramétricos

- Asumen que los datos tienen una determinada distribución
- El contraste es sobre alguno de los parámetros de una distribución
- Ejemplo: Test de la t de Student para una muestra

Contrastes No Paramétricos

- No asumen ninguna distribución para los datos
- En principio, son más flexibles

Entonces, ¿cuál usamos?

- Paramétricos, si se cumplen las hipótesis sobre los datos
- No paramétricos, en otro caso
- **OJO**: ¡¡Param./No param. no contrastan exactamente lo mismo!!

No olvidar de todas maneras que la primera parte del análisis de unos datos siempre involucra la descripción numérica y gráfica de los mismos.

Ejemplo

```
library(foreign)
ambiente<-read.spss(file="ambiente.sav", to.data.frame=TRUE)
attach(ambiente)
# Análisis descriptivo numérico
summary(ambiente)
by(OZONO,OZONO,length) # Nº de lugares clasf. por ozono
by(SULFATO, OZONO, mean) # Media de sulfato por grupo de ozono
by(PH, PROVIN, summary) # Est. resumen de PH por provincia
# Diagrama de cajas por factores
boxplot(SULFATO~PROVIN)
boxplot(PH~OZONO)
# Gráficos
hist(SULFATO, main="Histograma del SULFATO")
boxplot(PH, main="Diagrama de cajas del PH")
#Gráficos por grupos
par(mfrow=c(2,2))
hist(PH, main="Histograma del PH")
by(PH, PROVIN, function(X, xlim){hist(X, xlim=xlim)},xlim=range(PH))
```


Sección 2 | Modelos básicos con datos normales

Modelos básicos con datos normales

- Aunque es posible realizar análisis estadísticos y numéricos de gran complicación, no tiene ningún sentido el llevar a cabo un análisis más complicado del estrictamente necesario.
- De acuerdo con el principio de la navaja de Occam, la elección un modelo estadístico debe ser siempre lo más simple posible.
- Vamos a describir los contrastes de hipótesis más básicos que permiten resolver algunos de los análisis de datos más habituales que involucran una y dos muestras cuando los datos son normales.
- En concreto:
  - Problemas de una muestra (interés sólo en la variable respuesta)
    - Media
  - Problemas de dos muestras (interés en la variable respuesta pero explicada por un único factor de dos niveles)
    - Comparación de varianzas
    - Comparación de medias:
      muestras independientes y muestras emparejadas.

Inferencia en problemas de una muestra

Inferencia sobre la media de una población

Sea $X_1, \ldots, X_n$ una m.a. de una población $N(\mu, \sigma^2)$

- Si el interés es realizar inferencia sobre μ, podemos estimarla con $\bar{x}$.
- Además cuando σ^2 es desconocida, podemos hacer un contraste sobre μ del tipo $\begin{cases} H_0: & \mu \leq \mu_0 \\ H_A: & \mu > \mu_0 \end{cases}$ utilizando como regla de decisión:

  $$\text{“Rechazar } H_0 \text{ si } t_s = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} > t_{1-\alpha}(n-1)\text{”}$$

  o bien obteniendo su correspondiente p-valor.
- Si σ^2 es conocida podemos contrastar dicho contraste rechazando si $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > Z_{1-\alpha}$, o utilizando su correspondiente p-valor.

GEeitEm

Contraste unilateral de una media

Inferencia sobre la media de una población

Sea $X_1, \ldots, X_n$ una m.a. de una población $\mathrm{N}(\mu, \sigma^2)$

- Análogamente para el contraste $\begin{cases} \mathrm{H}_0: & \mu \geq \mu_0 \\ \mathrm{H}_A: & \mu < \mu_0 \end{cases}$ con σ^2 desconocida, podemos utilizar la regla:

  "Rechazar H_0 si $t_s = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} < t_\alpha(n-1)$",

  o el correspondiente p-valor.
- Si σ^2 es conocida se rechaza si $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} < Z_\alpha$, o bien se utiliza el p-valor.

Contraste bilateral de una media

Contrastes sobre la media de una población

Sea $X_1, \ldots, X_n$ una m.a. de una población $\mathrm{N}(\mu, \sigma^2)$

- Para el contraste bilateral $\begin{cases} \mathrm{H}_0: & \mu = \mu_0 \\ \mathrm{H}_A: & \mu \neq \mu_0 \end{cases}$ con σ^2 desconocida, la regla de decisión es:

  "Rechazar H_0 si $t_s > t_{1-\alpha/2}(n-1)$ ó $t_s < t_{\alpha/2}(n-1)$",

  o bien utilizar el p-valor.
- Si σ^2 es conocida se rechaza si $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} < Z_{\alpha/2}$ ó $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > Z_{1-\alpha/2}$, o bien se utiliza el p-valor.

Utilizando R para resolver contrastes de una media

Test paramétrico

El comando `t.test` nos permite realizar inferencia sobre los contrastes anteriores

```
t.test(x, y = NULL, alternative = c("two.sided", "less",
"greater"), mu = 0, paired = FALSE, var.equal = FALSE,
conf.level = 0.95, ...)
```

Ejemplo

```
t.test(PH, mu=4) # La media de PH distinta de 4?
t.test(SULFATO, alternative="less") # La media de SULFATO
es menor que 0?
t.test(SULFATO, mu=4, alternative="greater") # La media de
SULFATO es mayor que 4?
```


Análisis no param. de una muestra: Test de Wilcoxon

Descripción

- Contraste sobre la centralidad de una población (mediana)
- Observaciones independientes: $X_1, \ldots, X_n$
- Distribución simétrica de la población

Contraste

- H_0: **Mediana** $= \mu_0$
- H_A: **Mediana** $\neq \mu_0$

Ejemplo

```
x<-c(9,10,8,4,8,3,0,10,15,9)
wilcox.test(x, mu=5) #¿Es la mediana 5?
```

Ejemplo

En un estudio del crecimiento de plantas, un fisiólogo plantó 13 plantas de soja y les midió la altura al cabo de 16 días con la intención de comprobar si el crecimiento medio era superior a 20 cm.
Si los resultados fueron:
20.2, 22.9, 22.3, 20, 19.4, 22, 22.1, 22, 21.9, 21.5, 19.7, 21.5, 20.9,
¿qué podemos concluir al respecto?

Inferencia en problemas de dos muestras

Contrastes sobre la varianza de dos poblaciones

Sea $X_{11}, \ldots, X_{1n_1}$ una m.a. de N(μ_1, σ_1^2) y $X_{21}, \ldots, X_{2n_2}$ una m.a. de N(μ_2, σ_2^2)

- Para el contraste bilateral $\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_A: & \sigma_1^2 \neq \sigma_2^2 \end{cases}$ la regla de decisión es:

  "Rechazar H_0 si $F_s = \frac{s_1^2}{s_2^2} > F_{1-\alpha/2}(n_1-1, n_2-1)$ ó $F_s < F_{\alpha/2}(n_1-1, n_2-1)$"

  o bien utilizar el p-valor.

- Si el contraste es unilateral $\begin{cases} H_0: & \sigma_1^2 \leq \sigma_2^2 \\ H_A: & \sigma_1^2 > \sigma_2^2 \end{cases}$ se rechaza si $F_s = \frac{s_1^2}{s_2^2} > F_{1-\alpha}(n_1-1, n_2-1)$ o utilizando el p-valor. Análogamente con el otro contraste unilateral.

Comando `var.test` de R

```
var.test(respuesta ~ factor)
var.test(x, y, ratio = 1, alternative = c("two.sided", "less",
"greater"), conf.level = 0.95, ...)
```

Ejemplo

Un horticultor desea evaluar un nuevo insecticida que, según la publicidad, reduce los daños causados por los insectos. Con esta finalidad, realiza el siguiente experimento con 57 de los árboles de su huerto: trata 22 con el nuevo insecticida y los otros 35 con el antiguo. De los datos de la cosecha (en kilos) de estos árboles se obtuvieron los siguientes estadísticos:

	Nuevo	Antiguo
Media	249	233
Desviación estándar	39	45

Como paso previo para conocer si realmente es mejor el nuevo insecticida, se necesita comprobar si las varianzas de los dos grupos son o no son iguales. ¿Qué podemos concluir?

Comparación de las medias de dos muestras independientes

Contrastes sobre la media de dos poblaciones

Sea $X_{11},\ldots,X_{1n_1}$ una m.a. de $N(\mu_1,\sigma_1^2)$ y $X_{21},\ldots,X_{2n_2}$ una m.a. de $N(\mu_2,\sigma_2^2)$

- Si las varianzas son **iguales y desconocidas**, para el contraste bilateral $\left\{\begin{array}{ll} H_0: & \mu_1=\mu_2 \\ H_A: & \mu_1\neq\mu_2 \end{array}\right.$ la regla de decisión es:

  "Rechazar H_0 si $t_s > t_{1-\alpha/2}(n_1+n_2-2)$ ó $t_s < t_{\alpha/2}(n_1+n_2-2)$",

  donde $t_s = \frac{\bar{x}_1-\bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ y $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ es la estimación de la Varianza común. Alternativamente se puede utilizar el p-valor.

- Si el contraste es unilateral, $\left\{\begin{array}{ll} H_0: & \mu_1\leq\mu_2 \\ H_A: & \mu_1>\mu_2 \end{array}\right.$ se rechaza H_0 si el estadístico del contraste $t_s > t_{1-\alpha}(n_1+n_2-2)$ (análogamente con el otro contraste unilateral). Alternativamente se puede utilizar el p-valor.

Utilizando R para realizar tests *t*

Comando `t.test` de R

El comando anterior `t.test` se utiliza ahora dando valor a `y`

```
t.test(x, y = NULL, alternative = c("two.sided", "less",
"greater"), mu = 0, paired = FALSE, var.equal = FALSE,
conf.level = 0.95, ...)
```

pero se puede utilizar también

```
t.test(respuesta ~ factor)
```

Ejemplo

```
x <- c(0.80,0.83,1.89,1.04,1.45,1.38,1.91,1.64,0.73,1.46)
y <- c(1.15,0.88,0.90,0.74,1.21)
var.test(x,y); t.test(x, y, alternative="greater")
```


Comparación de 2 medias no param.: Test Mann-Whitney

Descripción

- Comparamos si dos poblaciones tienen la misma mediana
- Muestras: $\mathbf{x_1} = (x_1^1, \ldots, x_{n_1}^1)$ y $\mathbf{x_2} = (x_1^2, \ldots, x_{n_2}^2)$

Contraste

- H_0: **Mediana**$_1$ $=$ **Mediana**$_2$
- H_A: **Mediana**$_1$ $\neq$ **Mediana**$_2$

Ejemplo

```
x <- c(0.80,0.83,1.89,1.04,1.45,1.38,1.91,1.64,0.73,1.46)
y <- c(1.15,0.88,0.90,0.74,1.21)
wilcox.test(x, y) # opcional, alternative = "greater"
```

Ejemplo

Un horticultor desea evaluar un nuevo insecticida que, según la publicidad, reduce los daños causados por los insectos. Con esta finalidad, realiza el siguiente experimento con 57 de los árboles de su huerto: trata 22 con el nuevo insecticida y los otros 35 con el antiguo. De los datos de la cosecha (en kilos) de estos árboles se obtuvieron los siguientes estadísticos:

	Nuevo	Antiguo
Media	249	233
Desviación estándar	39	45

¿Es realmente mejor el nuevo insecticida?

Comparación no paramétrica de dos muestras emparejadas

Contrastes sobre la media de dos poblaciones

Sea $X_1, \ldots, X_n$ una m.a. de $N(\mu_1, \sigma_1^2)$ e $Y_1, \ldots, Y_n$ una m.a. de $N(\mu_2, \sigma_2^2)$

- Los datos vienen dados como pares $(X_1, Y_1), \ldots, (X_n, Y_n)$ porque entre ellos existe una relación de dependencia. Si definimos $D = X - Y$, tendremos una nueva muestra formada por las diferencias que tendrá su media ($\bar{d}$) y desviación estándar (s_d).
- Así, para resolver el contraste $\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_A: & \mu_1 \neq \mu_2 \end{cases}$ utilizaremos el contraste $\begin{cases} H_0: & \mu_d = 0 \\ H_A: & \mu_d \neq 0 \end{cases}$ cuya resolución es similar a la de los contrastes de una muestra (análogamente también los unilaterales).

`t.test` tiene una opción para muestras emparejadas

```
x <- c(1.83, 0.50,1.62,2.48,1.68,1.88,1.55,3.06,1.30)
y <- c(0.878,0.647,0.598,2.05,1.06,1.29,1.06,3.14,1.29)
t.test(x, y, paired=TRUE, alternative="greater")
```

Comparación de las medias de dos muestras emparejadas

Versión no paramétrica : Test de los signos

- Comparamos si dos poblaciones tienen la misma distribución
- Utilizamos el Test de Wilcoxon para una muestra con $\boldsymbol{y} = \boldsymbol{x_1} - \boldsymbol{x_2}$ y $\mu_0 = 0$

Contraste

- H_0: $\textbf{Mediana}_y = 0$
- H_A: $\textbf{Mediana}_y \neq 0$

Ejemplo

```
wilcox.test(x, y, paired = TRUE, alternative = "greater")
```


Ejemplo

Un proceso habitual en las industrias conserveras consiste en tratar las verduras con agua hirviendo antes de enlatar las mismas. El problema radica en la gran pérdida de vitaminas que sufren las verduras así tratadas. Se piensa que un método consistente en un lavado previo de las verduras con vapor de agua puede evitar la pérdida de vitaminas. Para comparar los dos métodos, se analizaron 10 grupos de judías provenientes de granjas diferentes. La mitad de las judías de un grupo se trataron con agua hirviendo y la otra mitad con vapor de agua. Se midió el contenido vitamínico de cada una de las mitades después del lavado, y se obtuvieron los siguientes resultados:

Grupo	1	2	3	4	5	6	7	8	9	10
Vapor	35	48	65	33	61	54	49	37	58	65
Agua	33	40	55	41	62	54	40	35	59	56

¿Se puede decir que realmente el método lavado con vapor de agua da mejores resultados que el del agua hirviendo?

Sección 3 | Análisis de la Varianza

Motivación: Comparaciones entre grupos

Muchos estudios aplicados se basan en la idea de comparar la media de varios grupos: unos que han recibido diferentes tratamientos y otro que no (y que recibe el nombre de control). Observar que tenemos una situación similar a la comparativa de la media de dos grupos (que resolvíamos con un test t), salvo que ahora tenemos más de dos grupos.

Ejemplos

Se pretende valorar la emisión de un cierto contaminante de 5 vehículos distintos, en concreto si el valor medio es similar en los cinco vehículos. Los datos siguientes son una parte de los obtenidos:

	Nissan	Fiat	Volkswagen	Jeep	Land Rover
	159.7	179.6	167.4	173.5	172.3
	161.5	173.9	163.0	182.4	168.9

Un agricultor posee tres campos en los que quiere comprobar el comportamiento de tres sistemas de regadío. Para ello, divide cada campo en tres zonas en cada una de las cuales aplica cada sistema. En el cuadro siguiente aparecen las cantidades (en kilos) recogidas en varios árboles clasificadas por campo y sistema utilizado:

	Sistema A	Sistema B	Sistema C
Campo 1	41, 37, 46, 32, 39, 40	51, 43, 53, 37, 56, 53	27, 29, 33, 17, 23, 28
Campo 2	28, 37, 42, 29, 35, 23	39, 43, 48, 51, 53, 47	19, 23, 28, 23, 31, 18
Campo 3	47, 56, 51, 43, 46, 44	49, 53, 51, 60, 52, 49	31, 33, 29, 27, 26, 32

Análisis de la Varianza: ANOVA.

- Una variable respuesta se puede modelizada en función de un conjunto de variables explicativas continuas o discretas.
- La situación más básica es la que sólo tenemos una variable explicativa (o factor de clasificación).
- Nuestro objetivo será pues valorar si existen diferencias en los valores de la variable respuesta en las diferentes categorías del factor de clasificación.
- El factor puede ser de:
  - Efectos fijos: cuando estemos interesados en las diferencias entre los grupos definidos por los niveles del factor. Nos interesa el efecto de cada tratamiento concreto en comparación con los otros.
  - Efectos aleatorios: No estamos interesados en las diferencias concretas entre los grupos definidos por los niveles del factor, es decir, no nos interesa el efecto de cada tratamiento concreto en comparación con los otros. Nos interesa la variabilidad entre grupos. Los diferentes niveles del factor, se consideran como una muestra aleatoria de todos los niveles que el investigador podría seleccionar.
- Diremos que el modelo (o el diseño) es equilibrado cuando el número de observaciones por nivel del factor sea el mismo.

ANOVA de un factor de efectos fijos

$$Y_{ij} = \underbrace{\mu + \alpha_i}_{\text{Comp. Sist.}} + \underbrace{\varepsilon_{ij}}_{\text{Comp. Aleat.}} \quad , \quad i = 1, \ldots, a\,; \quad j = 1, \ldots, n_i$$

- $\sum_{i=1}^{a} \alpha_i = 0$ para evitar la sobreparametrización. Equivalente a $Y_{ij} = \mu_i + \varepsilon_{ij}$.
- μ representa la media global de la población
- α_i es la desviación de la media del grupo i de la media global
- ε_{ij}: desviación del individuo j de la media del grupo i; $\varepsilon_{ij} \sim N(0, \sigma^2)$ independientes

ANOVA de un factor de efectos aleatorios

$$Y_{ij} = \underbrace{\mu + A_i}_{\text{Comp. Sist.}} + \underbrace{\varepsilon_{ij}}_{\text{Comp. Aleat.}} \quad , \quad i = 1, \ldots, a\,; \quad j = 1, \ldots, n_i$$

- $A_i \sim N(0, \sigma_A^2)$
- μ representa la media global de la población
- A_i es el efecto aleatorio debido a ser del grupo i
- ε_{ij}: desviación del individuo j de la media del grupo i; $\varepsilon_{ij} \sim N(0, \sigma^2)$ independientes

Descomposición suma de cuadrados y Tabla ANOVA

- Si denotamos con $\overline{y}_{i.}$ a la media de cada grupo, y con $\overline{y}_{..}$ a la media global de todos los datos ($= \frac{\sum_{i,j} y_{ij}}{N} = \frac{\sum_i n_i \overline{y}_{i.}}{N}$ con $N = \sum_i n_i$), podemos descomponer cada dato de la siguiente manera:

$$y_{ij} - \overline{y}_{..} = (y_{ij} - \overline{y}_{i.}) + (\overline{y}_{i.} - \overline{y}_{..})\,,\, i = 1, \ldots, a\,; j = 1, \ldots, n_i$$

- Elevando al cuadrado y sumando para todos los valores de i y de j:

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{..})^2}_{\text{SS Total}} = \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i.})^2}_{\text{SS Intra o Error}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\overline{y}_{i.} - \overline{y}_{..})^2}_{\text{SS Entre}}$$

- En forma de Tabla de ANOVA (Análisis de la Varianza):

F. Variación	SS	gl	MS
Entre	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\overline{y}_{i.} - \overline{y}_{..})^2$	$a-1$	MS Entre = SS Entre/(a-1)
Error	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i.})^2$	$N-a$	MS Error = SS Error/(N-a)
Total	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{..})^2$	$N-1$	

- ¡Las sumas de cuadrados se reescriben para que los cálculos sean más cómodos!

Inferencia sobre los parámetros

ANOVA de un factor de efectos fijos

- Estimación parámetros (por máxima verosimilitud):
  1. $\hat{\mu} = \overline{y}_{..}$
  2. $\hat{\alpha}_i = \overline{y}_{i.} - \overline{y}_{..}$
  3. $\hat{\sigma}^2 =$ MS Error
- Contraste de hipótesis: el cociente $F_s = \frac{\text{MS Entre}}{\text{MS Error}}$ es el estadístico de contraste de

$$\left.\begin{array}{ll} H_0: & \mu_1 = \mu_2 = \ldots = \mu_a \\ H_A: & \text{no } H_0 \end{array}\right\} \equiv \left.\begin{array}{ll} H_0: & \alpha_1 = \alpha_2 = \ldots = \alpha_a = 0 \\ H_A: & \text{no } H_0 \end{array}\right\}$$

  y su distribución en el muestreo bajo H_0 es una $F(a-1, N-a)$.

- Observar que si $a = 2$ el contraste resultante equivale a una comparación de dos muestras independientes. Se puede demostrar que ambos estadísticos de contraste (el F_s y el T_s) están relacionados (en concreto: $F_s = T_s^2$), y por tanto llegan a la misma conclusión.

Inferencia sobre los parámetros (2)

ANOVA de un factor de efectos aleatorios

- Estimación parámetros (por máxima verosimilitud):
  1. $\hat{\mu} = \bar{y}_{..}$;
  2. $\hat{\sigma}^2 =$ MS Error
  3. $\hat{\sigma}^2_A = \frac{\text{MS Entre - MS Error}}{n_0}$ (si esta cantidad es menor que 0, $\hat{\sigma}^2_A = 0$) con:
     - $n_0 = \frac{1}{N(a-1)}\left(N^2 - \sum_{i=1}^{a} n_i^2\right)$.
     - Si el diseño es equilibrado, $n_0 = n_i = n$.
- Contraste de hipótesis: el cociente $F_s = \frac{\text{MS Entre}}{\text{MS Error}}$ es el estadístico de contraste de

$$\left.\begin{array}{ll} H_0: & \sigma^2_A = 0 \\ H_A: & \sigma^2_A \geq 0 \end{array}\right\}$$

  y su distribución en el muestreo bajo H_0 es una $F(a-1, N-a)$.
- Así pues, mismo estadístico de contraste pero diferentes hipótesis, ya que el modelo es diferente y las conclusiones también son diferentes.

Validez del modelo: condiciones de aplicabilidad

- La inferencia en ambos casos está basada en tres condiciones que vienen determinadas por los propios modelos.
- $\varepsilon_{ij} \sim N(0, \sigma^2)$ independientes
  - *Homocesdasticidad*: misma varianza en los grupos analizados
  - *Normalidad*: de los datos de cada grupo
  - *Independencia*: tanto de las muestras entre sí como de las observaciones en cada grupo
- Tenemos dos opciones para comprobar estas condiciones:
  - Antes del ANOVA, utilizar tests de homogeneidad de varianzas (Bartlett, Fligner-Killeen o Levene), tests de normalidad (Kolmogorov-Smirnov o Shapiro-Wilks) y tests de independencia (rachas), tal y como hacíamos en la comparación de dos medias
  - Después del ANOVA, analizar los residuos resultantes del ajuste, tal y como hacíamos en regresión
- Si no se cumplen las condiciones:
  - podemos transformar los datos (p. e. utilizando Box-Cox)
  - utilizar un método no paramétrico: el test de Kruskal-Wallis, una generalización del test de Mann-Whitney

Tarea

Los siguientes datos provienen de un experimento realizado en la estación experimental de Rothamsted. El objetivo era medir la eficacia de tres insecticidas, el clorodinitrobenceno (CN), el carbón disulfido (CD) y un preparado propio denominado cymag (CM). Cada insecticida se aplicó a dosis normal (1) y doble (2). Por último se contó con un grupo control al que no se aplicó ningún insecticida. Los pesticidas se aplicaron antes de la siembra del trigo, y los datos recogidos muestran el incremento del número de gusanos encontrados en cada parcela después de la recolección del trigo.

	Insecticida					
Control	1CN	1CD	1CM	2CN	2CD	2CM
466	222	194	306	92	166	28
421	219	221	176	114	172	179
561	332	308	215	80	111	165
433	298	256	199	128	80	82

1. Especifica un modelo estadístico adecuado para analizar este experimento y explica el significado de sus parámetros.

Tarea

2. Comenta las hipótesis de aplicabilidad que deberían cumplirse para poder analizar estos datos con la técnica especificada en el apartado anterior.
3. ¿Existen diferencias estadísticamente significativas en el incremento de gusanos dependiendo del insecticida aplicado? En otras palabras, ¿hay efecto insecticida? Plantea y resuelve el contraste adecuado. Ayuda: comprueba para ello que con esos datos se obtiene la siguiente tabla ANOVA:

Fuentes de variación	SS	gl	MS	F
Insecticida	392447	6	65408	22.473
Residual	61121	21	2911	

4. ¿Cual es el alcance de las conclusiones que te aporta el contraste que has realizado?

Sección 4 Comparaciones

Comparaciones a posteriori

- Cuando se rechaza la hipótesis nula en un contraste en el que el interés es las diferencias entre grupos (efectos fijos), la conclusión es que al menos uno de los grupos tiene la media diferente, es decir, que hay un efecto de ese factor que estamos analizando.
- Pero no nos aclara ni que grupo es el que tiene la media diferente ni si es uno sólo.
- Una posibilidad para mejorar las conclusiones es utilizar comparaciones dos a dos de las medias de los grupos y en algunas de ellas hay que tener presente la protección del error global:
  - Método de Tukey
  - Método basado en las diferencias significativas utilizando la corrección de Bonferroni,
  - y otros muchos más (Scheffé, Student-Newman-Keuls, Dunn-Sidak, etc.)

Comparaciones a posteriori: caso particular diseño equilibrado (n tamaño grupo)

1. Cálculo y ordenación de las medias de los grupos.
2. Cálculo de las diferencias dos a dos de las medias ordenadas y construcción de una tabla de diferencias de medias.
3. Obtención de las diferencias significativas (en base a cualquiera de los métodos) como aquellas que superan los valores:
   - Método Bonferroni. Si α' es el error global máximo a cometer:

     $$LSD_\alpha = \sqrt{F_{\alpha(1,N-a)}}\sqrt{\frac{2}{n}MSError}$$

     con $\alpha = \frac{\alpha'}{k}$ siendo k el número de comparaciones dos a dos.
   - Método Tukey:

     $$MSR_\alpha = Q_{\alpha,(a,N-a)}\sqrt{\frac{MSError}{n}}$$

     con $Q_{\alpha,(a,N-a)}$ valor crítico en la tabla de rangos estudentizados.
4. Obtención de subgrupos homogéneos: aquellos subgrupos con medias similares.

Tarea

1. Obtener los grupos homogéneos resultantes de aplicar el método de Tukey con los datos de los insecticidas. ¿De qué manera observas que se amplian las conclusiones que puedes aportar sobre el problema planteado?
2. Para controlar el posible impacto medioambiental que supondría el incendio de varias fábricas de tejidos próximas a un bosque, se determinó el tiempo (en segundos) que tardaban en arder 5 vestidos, elegidos al azar, realizados en cada una de ellas. Del análisis de los datos se obtuvieron los siguientes resultados:

Fábrica	1	2	3	4	5
Media	16.78	11.76	10.24	11.98	15.26
Desv. Típica	1.167	2.3298	1.1437	1.862	0.9182

   1. Si medimos la peligrosidad de una fábrica por el tiempo que tardan en arder sus vestidos, ¿hay evidencia para pensar que los tejidos de las fábricas influyen en su peligrosidad?
   2. Calcula los subgrupos homogéneos resultantes de aplicar tanto el método de la diferencia significativa con la corrección de Bonferroni como el método de Tukey. ¿Qué conclusiones puedes extraer de estos grupos homogéneos? ¿Cuáles son tus conclusiones globales sobre el análisis realizado?

Comparaciones previamente planificadas (o a priori)

- Existe otra posibilidad de completar la información que nos aporta un contraste de ANOVA de un factor de efectos fijos.
- Consiste en resolver contrastes que previamente son de interés (a diferencia de las comparaciones a posteriori donde no había un interés planificado, si no que eran los propios datos los que nos aportaban los grupos homogéneos).
- Existen dos posibilidades:
  - Comparar k de los a grupos
  - Comparar una combinación lineal de las medias (contrastes de medias)

Comparación de k de las a medias: procedimiento

Para contrastar

$$\left.\begin{array}{ll} H_0: & \mu_{(1)} = \mu_{(2)} = \ldots = \mu_{(k)} \\ H_A: & \text{no } H_0 \end{array}\right\} \equiv \left.\begin{array}{ll} H_0: & \alpha_{(1)} = \alpha_{(2)} = \ldots = \alpha_{(k)} = 0 \\ H_A: & \text{no } H_0 \end{array}\right\}$$

utilizamos el estadístico de contraste $F_s = \frac{\text{MS Entre } k \text{ grupos}}{\text{MS Error}}$ que se distribuye bajo H_0 como una $F(k-1, N-a)$ siendo MS Entre k grupos $= \frac{\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{k-1}$.

Contrastes de medias: procedimiento

1. Determinar el *contraste de medias* que se quiere realizar. Puede ser cualquier comparación a priori del tipo:

$$\left.\begin{array}{ll} H_0: & c_1\mu_1 + c_2\mu_2 + \ldots + c_a\mu_a = 0 \\ H_A: & \text{no } H_0 \end{array}\right\} \equiv \left.\begin{array}{ll} H_0: & \sum_{i=1}^{a} c_i\mu_i = 0 \\ H_A: & \text{no } H_0 \end{array}\right\}$$

   que cumpla que los coeficientes $\sum_i c_i = 0$.

2. Calcular el estadístico de contraste

$$F_s = \frac{\text{SS Entre}}{\text{MS Error}}$$

   que se distribuye bajo H_0 como una $F(1, N-a)$ siendo

$$\text{SS Entre} = \frac{\left(\sum_{i=1}^{a} c_i \bar{y}_{i.}\right)^2}{\sum_{i=1}^{a} \frac{c_i^2}{n_i}}$$

3. Rechazar la hipótesis nula de acuerdo al p-valor asociado al estadístico anterior.

Contrastes de medias: ortogonalidad

- Dos *contrastes de medias*

$$\left.\begin{array}{ll} H_0: & \sum_{i=1}^{a} c_i\mu_i = 0 \\ H_A: & \text{no } H_0 \end{array}\right\} \quad \text{y} \quad \left.\begin{array}{ll} H_0: & \sum_{i=1}^{a} d_i\mu_i = 0 \\ H_A: & \text{no } H_0 \end{array}\right\}$$

  son ortogonales cuando $\sum_{i=1}^{a} \frac{c_i d_i}{n_i} = 0$:
- Cuando dos contrastes de medias son ortogonales, las dos comparaciones son independientes y aportan información complementaria.
- Es posible descomponer una comparación de a grupos en $a-1$ contrastes de medias ortogonales entre sí (cada uno con un grado de libertad).
- Las comparaciones de k de las a medias también se pueden descomponer en $k-1$ contrastes de medias ortogonales entre sí.
- Además, la SS Entre de la comparación de a grupos también se puede descomponer como la SS Entre de cada contraste de medias.

Tarea

En un estudio sobre el comportamiento de un parásito, se visitaron cinco zonas de interés situadas en la Comunitat Valenciana (Alcublas, Jarafuel, Jalance, Enguera y el Saler; las cuatro primeras localizadas en el interior de la provincia de Valencia y la última en la costa). En cada una de las zonas se analizaron varios parásitos. En concreto se estudió su peso (en gramos), siendo el objetivo comprobar si existían diferencias en el peso en las diferentes zonas consideradas.

	Alcublas	Jarafuel	Jalance	Enguera	el Saler
Media	2.40	2.37	2.45	2.07	1.47
Varianza	0.48	0.39	0.21	0.32	0.46
Tamaño muestral	200	200	200	200	200

1. Especifica un modelo estadístico adecuado para analizar este experimento y explica el significado de sus parámetros.
2. Valora las hipótesis de aplicabilidad que deberían cumplirse para poder analizar estos datos con la técnica especificada en el apartado anterior.
3. ¿Existen diferencias estadísticamente significativas en el peso de los parásitos en las cinco zonas analizadas?
4. ¿Como analizarías la hipótesis de que en la zona costera el peso difiere de la zonas más interiores? ¿Puedes hacerlo con los datos disponibles?

Sección 5 | ANOVA de un factor con R

ANOVA de un factor con R

El objetivo de esta sección es trabajar la resolución de los problemas de ANOVA de un factor utilizando `R`.

- Aunque podemos utilizar el comando `lm` (`R` toma como referencia un nivel del factor y utiliza variables *dummy* para el resto de niveles del factor),
- habitualmente se utiliza el comando `aov`.

Algunos elementos importantes

- La función `summary` sobre un objeto tipo `aov` produce la tabla de ANOVA necesaria para realizar el contraste de comparación de las medias.
- `fitted.values`: Valores ajustados, $\hat{y}_i$
- `residuals`: Valores de los residuos (no tipificados)
- Con el comando `rstandard` accedemos a los residuos estandarizados.

Validación del ajuste realizado

- Podemos realizar la comprobación de las condiciones de aplicabilidad antes de realizar el ANOVA:
  1. *Homocesdasticidad*: para comprobar si las varianzas son homogéneas podemos utilizar el test de Bartlett (`bartlett.test()`), el de Fligner-Killeen (`fligner.test()`) o el de Levene (`leveneTest()` de la librería `car`).
  2. *Normalidad*: para comprobarla podemos utilizar el test de Kolmogorov Smirnov (`ks.test()`) o el de Shapiro Wilk (`shapiro.test()`).
  3. *Indepedencia*: comprobando si existen rachas (grupos de valores ininterrumpidos en la misma dirección) podemos valorar la independencia de los datos.
- Como en regresión con el comando `plot(objetoaov)` tenemos cuatro gráficas que nos permiten validar la adecuación del modelo.
- Si no tenemos condiciones de aplicabilidad podemos transformar los datos (la mejor transformación nos la da la función `boxcox`) o utilizar el test de Kruskal Wallis (`kruskal.test`) que contrasta la hipótesis:

$$H_0\text{: Mediana}_1 = \ldots = \text{Mediana}_p$$

Comparaciones a priori: contrastes de medias

- Si tenemos un contraste de medias:

$$\left.\begin{array}{ll} H_0: & c_1\mu_1 + c_2\mu_2 + \ldots + c_a\mu_a = 0 \\ H_A: & \text{no } H_0 \end{array}\right\} \equiv \left.\begin{array}{ll} H_0: & \sum_{i=1}^{a} c_i\mu_i = 0 \\ H_A: & \text{no } H_0 \end{array}\right\}$$

  con $\sum_i c_i = 0$, podemos resolverlo utilizando `R` utilizando el hecho de que las variables de tipo `factor` tienen entre sus propiedades los contrastes en que descomponer su efecto.
  Por defecto, contrasta si la media de cada nivel coincide con la media del nivel de referencia.
- Para ello utilizamos el comando `contrasts()` construyendo la matriz que define los contrastes, y luego asociándola al factor.

Comparaciones a priori: contrastes de medias

Por ejemplo, si queremos contrastar si el grupo 1 es similar al resto de cuatro grupos:

```
m.contrs ← matrix(c(-4,rep(1,4)),ncol=1)
contrasts(factor) ← m.contrs
ajuste.contrs ← aov(var.respuesta ~ factor)
summary.aov(ajuste.contrs,
            split=list(factor=list("G1.vs.resto"=1)))
```

Observar que la instrucción `summary.aov()` actuando sobre el objeto resultante de un ajuste mediante `aov()` produce una tabla con la valoración de los efectos del factor que, al corresponder con los contrastes que le hemos asociado, responden a las preguntas formuladas.

Si quisiéramos contrastar además que el segundo grupo es similar al tercero y al cuarto:

```
m.contrs ← matrix(c(-4,rep(1,4),c(0,-2,1,1,0)),ncol=2)
contrasts(factor) ← m.contrs
ajuste.contrs ← aov(var.respuesta ~ factor)
summary.aov(ajuste.contrs,
   split=list(factor=list("G1.vs.resto"=1,"G2.vs.G3yG4"=2)))
```


Comparaciones a posteriori

Para resolver comparaciones a posteriori utilizando R podemos utilizar las funciones:

- `TukeyHSD`, que nos da la diferencia entre todas las medias dos a dos, y un intervalo de confianza para dicha diferencia junto a un p-valor ya ajustado que nos indica si dicha diferencia es significativa. A partir de esta información es muy sencillo construir los subgrupos homogéneos como aquellos que contienen grupos con medias cuya diferencia no sea significativa.
- `LSD.test`, `SNK.test`, `scheffe.test` son tres funciones de la librería `agricolae` que nos indican los subgrupos homogéneos resultantes de aplicar dichos métodos, junto con las cotas a partir de las cuales las diferencias son significativas.

¿Cambia la longitud de la sección de un guisante con la adición de azúcares?

Situación

Los siguientes datos provienen del ejemplo expuesto en el libro de Sokal y Rohlf (Biometría: principios y métodos estadísticos en la investigación biológica) y hace referencia a un estudio sobre la longitud de secciones de guisantes (en unidades oculares de 0.114 mm.) criados en cultivos con adición de distintos azúcares.

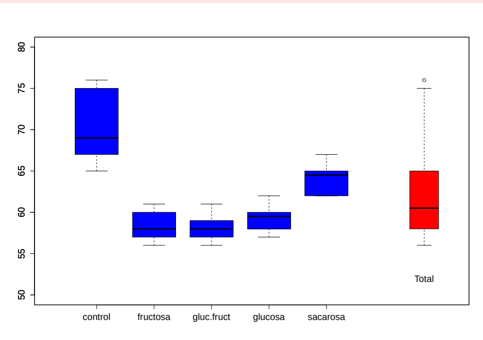
control	glucosa	fructosa	gluc+fruct	sacarosa
75	57	58	58	62
67	58	61	59	66
70	60	56	58	65
75	59	58	61	63
65	62	57	57	64
71	60	56	56	62
67	60	61	58	65
67	57	60	57	65
76	59	57	57	62
68	61	58	59	67

Describe el diseño que consideres más adecuado para valorar si hay cambios en la sección dependiendo del tipo de azúcar administrado.

Lectura y descriptiva: Código R

```
setwd("~/Documentos/Docencia/Bancos-Datos")
guis ← read.table(file='guisantes.dat', header=T)
attach(guis)
by(longitud,tratamiento,summary)
by(longitud,tratamiento,var)
boxplot(longitud~tratamiento,boxwex=0.75,
        ylim=c(50,80),xlim=c(0.5,7),col=4)
boxplot(longitud,add = TRUE, boxwex=1.5,at=6.5,col=2)
text(6.5,52,"Total")
```

Salidas R



Validación condiciones aplicabilidad antes: Código R

```
bartlett.test(longitud ~ tratamiento)
fligner.test(longitud ~ tratamiento)
library("car")
leveneTest(longitud ~ tratamiento)
by(longitud,tratamiento,shapiro.test)
runs.test(as.factor(longitud > median(longitud)))
```

Salidas R

```
Bartlett'sK-squared=13.9386,df=4,p-value=0.007494
Fligner-Killeen:medchi-squared=10.2645,df=4,p-value=0.0362
Levene-p-value=0.003468**

#Shapiro-Wilks-->normalidad

RunsTest
data:as.factor(longitud>median(longitud))
StandardNormal=-1.9258,p-value=0.0002032
```


Validación condiciones aplicabilidad después: Código R

```
guis.aov ← aov(longitud ~ tratamiento)
summary(guis.aov)
model.tables(guis.aov,type="means")
model.tables(guis.aov,type="effects")

# Tabla ANOVA
             Df  Sum Sq  Mean Sq  F value    Pr(>F)
tratamiento   4  1077.3   269.33    49.37  6.74e-16 ***
Residuals    45   245.5     5.46

Table of means: Grand mean  61.94

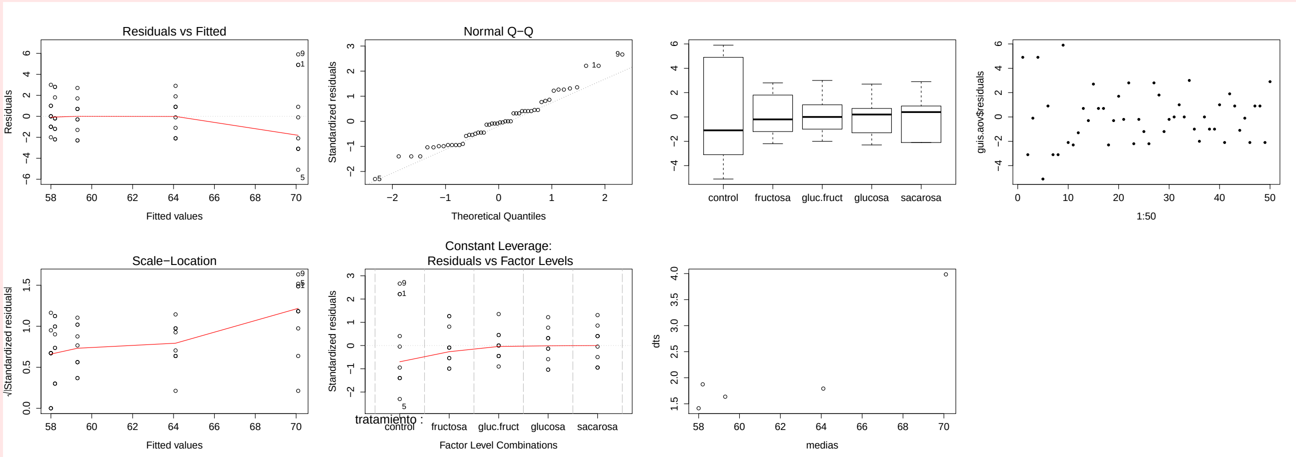
Table of means: tratamiento
   control   fructosa  gluc.fruct    glucosa   sacarosa
      70.1       58.2        58.0       59.3       64.1

Tables of effects tratamiento
   control   fructosa  gluc.fruct    glucosa   sacarosa
      8.16      -3.74       -3.94      -2.64       2.16
```

Validación condiciones aplicabilidad después: Código R

```
par(mfrow=c(2,2)); plot(guis.aov)
plot(guis.aov$model$tratamiento,guis.aov$residuals)
plot(1:50,guis.aov$residuals,pch=20)
medias ← tapply(longitud,tratamiento,mean)
dts ← tapply(longitud,tratamiento,sd)
plot(medias,dts)
```

Salidas R



Test no paramétrico y transformaciones: Código R

```
guis.ks ← kruskal.test(longitud,tratamiento)
guis.ks

#Kruskal-Wallis chi-squared=38.4368, df=4, p-value=9.105e-08

# transformaciones
library(MASS)
bc ← boxcox(guis.aov,lambda=seq(-6,0,length=20))
lambda ← bc$x[which.max(bc$y)]
lambda
# [1] -3.575758
trans.longitud ← longitud^(-3)

#tras transformar las condiciones aplicabilidad se cumplen
bartlett.test(trans.longitud,tratamiento)
by(trans.longitud,tratamiento,shapiro.test)
guis.trans.aov←aov(trans.longitud~tratamiento)
summary(guis.trans.aov)
model.tables(guis.trans.aov,type="means")
```

Los biólogos que diseñaron el experimento estaban interesados en el efecto de los distintos azúcares en el tamaño de los guisantes. Las preguntas más interesantes, formuladas en términos de comparaciones de grupos (a través de sus medias) eran:

1. ¿Influye cualquiera de los tipos de azúcar?, es decir, ¿difiere el grupo de control (sin azúcar) del resto de grupos reunidos como si fueran uno sólo?
2. ¿Existen diferencias entre la sacarosa y los otros dos tipos de azúcar?
3. ¿Hay diferencia entre utilizar glucosa o fructosa puras y o ambas combinadas?
4. ¿Difieren los grupos tratados con glucosa y fructosa?

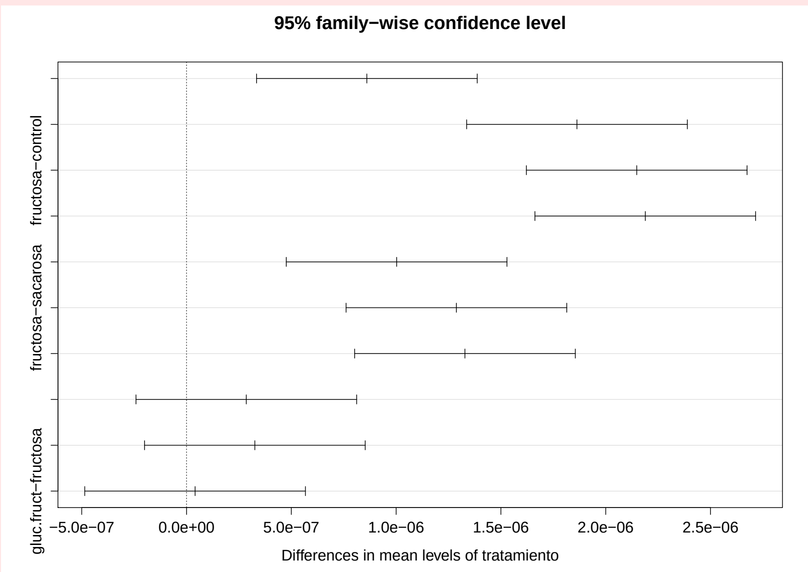
Comparaciones a priori: Código R

```
contrastes < -matrix(c(1,rep(-1/4,4),
                     0,rep(-1/3,3),1,
                     0,-1/2,1,-1/2,0,
                     0,-1,0,1,0),ncol=4)
contrasts(tratamiento) ← contrastes
guis.contrastes ← aov(trans.longitud~tratamiento)
summary.aov(guis.contrastes,
    split=list(tratamiento=list("Ctr.vs.azucs."=1,
                                "Sac.vs.restoazucs."=2,
                                "Gl.yFr.vs.ambascombs."=3,
                                "Gl.vsFr."= 4 )))
```


Comparaciones a posteriori: Código R

```
ajuste.tukey ← TukeyHSD(guis.trans.aov,ordered=T)
ajuste.tukey
plot(ajuste.tukey)
```

Salidas R



Comparaciones a posteriori: Código R

```
library("agricolae")
ajuste.LSD ← LSD.test(guis.trans.aov,"tratamiento",
              p.adj="bonferroni")
ajuste.LSD
```

Salidas R

```
$parameters
  Df ntr bonferroni
  45   5   2.952079
$groups
         trt        means M
1 gluc.fruct 5.141309e-06 a
2 fructosa   5.100475e-06 a
3 glucosa    4.815334e-06 a
4 sacarosa   3.812881e-06 b
5 control    2.952451e-06 c
```


Comparaciones a posteriori: Código R

```
ajuste.SNK ← SNK.test(guis.trans.aov,"tratamiento")
ajuste.SNK
ajuste.Scheffe ← scheffe.test(guis.trans.aov,"tratamiento")
ajuste.Scheffe
```

Salidas R

```
$groups
         trt        means M
1 gluc.fruct 5.141309e-06 a
2 fructosa   5.100475e-06 a
3 glucosa    4.815334e-06 a
4 sacarosa   3.812881e-06 b
5 control    2.952451e-06 c
```

Tarea: continuación problema incremento gusanos

1. Contrasta la posible igualdad de medias en el número de gusanos que proliferaron con los diferentes tratamientos, planteando el contraste de hipótesis adecuado.
2. A la vista del resultado obtenido, contrasta las posibles diferencias entre todos los grupos utilizando el método de Bonferroni y de Tukey. Construye los subgrupos homogéneos y comenta los resultados obtenidos.
3. Completa tus conclusiones considerando las siguientes comparaciones:
   - Una comparación entre el grupo control y los insecticidas
   - Una comparación entre los insecticidas a dosis 1 y los insecticidas a dosis 2
   - Una comparación entre los insecticidas a dosis 1
   - Una comparación entre los insecticidas a dosis 2

   Para cada una de estas comparaciones:
   - Descompón en contrastes ortogonales las comparaciones con más de 1 grado de libertad.
   - ¿Qué parejas de contrastes resultan ser ortogonales entre sí?
   - Resuelve los contrastes y comenta los resultados.
4. Valora las condiciones de aplicabilidad en este problema, y caso de que no se cumplieran, propón una transformación adecuada y valora entonces si existe el efecto que antes has estudiado.

Licencia de este material